DOCUMENT RESUME

ED 390 945 TM 024 600

AUTHOR Parkes, Jay; Suen, Hoi K.
TITLE The Preferability of Constrained Optimization in Determining the Number of Prompts, Modes of Discourse, and Raters in a Direct Writing Assessment.
PUB DATE Apr 95
NOTE 27p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Algorithms; *College Students; *Educational Assessment; *Generalizability Theory; Graduate Students; Higher Education; *Interrater Reliability; *Test Construction; *Writing Tests
IDENTIFIERS Branch and Bound Method; Direct Assessment

ABSTRACT

This study demonstrates the advantages of using a constrained optimization algorithm to explore the optimal number of prompts, modes of discourse, and raters for achieving an acceptable level of reliability during a direct writing assessment. Writing samples elicited from 50 college students were rated by 3 graduate students and the scores submitted to a generalizability analysis (G-study). The variance components estimated in the G-study were then used in a branch-and-bound integer programming procedure to determine the optimal number of raters, modes, and prompts to produce a reliable writing assessment. Four different scenarios were examined to show how the optimal answer changes based on the priorities of the measurement situation. (Contains 2 tables, 3 figures, and 24 references.) (Author/SLD)

The Preferability of Constrained Optimization in Determining the Number of Prompts, Modes of Discourse, and Raters in a Direct Writing Assessment

Jay Parkes and Hoi K. Suen

The Pennsylvania State University

Paper presented at the National Council on Measurement in Education Annual Meeting

San Francisco, CA: April, 1995

Abstract

This study demonstrates the advantages of using a constrained optimization algorithm to explore the optimal number of prompts, modes of discourse, and raters for achieving an acceptable level of reliability during a direct writing assessment. Writing samples elicited from college students were rated and the scores submitted to a generalizability analysis. The variance components estimated in the G-study were then used in a branch-and-bound integer programming procedure to determine the optimal number of raters, modes, and prompts to produce a reliable writing assessment. Four different scenarios are examined to show how the optimal answer changes based on the priorities of the measurement situation.

The Preferability of Constrained Optimization in Determining the Number of Prompts, Modes of Discourse, and Raters in a Direct Writing Assessment

The movement toward a direct assessment of writing over the last eighteen years has encountered many obstacles (Huot, 1990), several of which are ones that all performance assessments face (Linn, Baker, and Dunbar, 1991). Two of these are of particular interest in this study: generalizability and cost and efficiency (Linn et al., 1991).

In performance assessment, great strides have been made recently to tackle the problems associated with generalizability. Cost has also been a consideration. Sanders, Theunissen, and Baas (1992) have developed a framework for the simultaneous consideration of cost and generalizability in the development of assessments. Writing assessment, which is a special case of performance assessment, makes an appropriate testing ground for this new framework since the broad and complex nature of writing itself requires an equally broad and complex assessment of writing prowess (e.g. Raymond, 1982). Such an assessment, however, could be extremely expensive and time consuming (White, 1986). Therefore, the vexing question is this: How could an assessment be designed that would be broad and complex -- that is, generalizable -- and also be cost and time efficient. This study investigates the use of a branch-and-bound algorithm proposed by Sanders et al. (1992) to optimize the numbers of facets in a given assessment scenario to achieve pre-determined levels of reliability and cost.

The Problem of Generalizability

Generalizability is considered to be the size of the domain to which the findings from a measure can be applied. It is often expressed as the concurrence of the scores on two or more tasks or items. For instance, if a student writes an excellent descriptive

paragraph and an excellent narrative one, the evidence would support the conclusion that either score would generalize to writing ability. The determination might then be that, based on either one of these pieces of writing, the student is a good writer. However, if a student wrote an excellent descriptive paragraph but a poor narrative one, the evidence would be weak to support generalizing either paragraph to writing ability. That is, it would be inappropriate to say the student is a good (or poor) writer, but better to say that the student is a good descriptive writer but a poor narrative writer. This concurrence can be numerically expressed as a generalizability coefficient, which is obtained through the application of the framework of generalizability theory.

Generalizability theory (e.g. Brennan, 1992) is a framework for establishing the reliability of a measurement. One of its advantages over classical test theory is that it allows variance components from different facets of the measurement situation to be considered in an estimation of reliability. For example, in an essay test, there are questions, students' responses, scoring criteria, and raters, any of which can contribute to the variance in scores. Classical test theory can only address one at a time (i.e. inter-rater reliability or internal consistency). Generalizability theory can consider all of them simultaneously.

The problem of generalizability in performance assessments, and direct writing assessments in particular, is that of agreement among levels of a certain factor such as task, rater (to some degree), and others. These difficulties are exhibited in a variety of types of performance assessments by several concurrent lines of inquiry, including those: by Shavelson and colleagues [e.g. Shavelson and Baxter (1992), and Shavelson, Baxter and Gao (1993)]; from the Vermont Portfolio Assessment program [e.g. Koretz, Klein, McCaffrey, and Stecher (1994); Koretz, Stecher, Klein, McCaffrey, and Deibert (1994); and Koretz, Stecher, Klein, and McCaffrey (1994)]; and by McWilliam and Ware (1994).

Shavelson and colleagues have worked primarily with performance assessments in elementary level general science. By using the framework of generalizability theory, they

have been able to demonstrate that the greatest contributing facet to low reliability estimates is the task (e.g. Shavelson, Baxter and Gao, 1993). Furthermore, they project that by increasing the number of tasks, a more reliable assessment will result. Koretz and colleagues have worked with portfolio assessments of math and writing and, again, through generalizability theory, identified raters and tasks as sources of error variance (Koretz, Stecher, Klein, McCaffrey, and Deibert, 1994). They, too, explore the possibility of increasing the number of tasks and the number of raters to achieve a more acceptable estimate of reliability. McWilliam and Ware (1994) used generalizability theory to examine the assessment of young children's engagement, and identified the number of sessions or observations as being a large source of error variance. They estimated the minimum number of sessions that would be necessary to create an acceptably reliable assessment.

The findings from several studies have illustrated an analogous phenomenon in direct writing assessment (Prater and Padia, 1983; Quellmalz, Capell, and Chou, 1982; Engelhard, Gordon and Gabrielson, 1991; Kegley, 1986). The findings indicate that scores for the same writer will vary across different discourse modes , such as expressive and persuasive, which could create a generalizability problem if more than one mode were used. Many of these researchers come to the conclusion that single mode assessments should be used to gain reliable results, to which a warning to the consumer is to be attached stating the limited generalizability. Several others, for example Breland, Camp, Jones, Morris, and Rock (1987), disagree. Breland et al. recommended as one possibility that reliability problems in essay examinations be overcome by increasing the number of writing samples.

Generalizability theory has been proposed as the "natural framework" (Linn et al., 1991) in which to explore reliability issues in multidimensional assessments and will be utilized in the present study. Shavelson et al. (1992, 1993); Koretz et al. (1994); and McWilliam and Ware (1994) all used this framework, while the other researchers (Prater

and Padia, 1983; Quellmalz et al., 1982; Engelhard et al., 1991; Kegley, 1986; Breland et al., 1987) did not.

The problems of cost and efficiency

Linn et al. (1991) point out that performance assessments, being labor-intensive, are not as inexpensive as a multiple choice test. White (1986), however, holds that, when designed properly, a direct assessment of writing can be conducted with comparable expense to that of multiple choice assessment. This divergence notwithstanding, White (1986) recognized that the expenses are different for the two forms, the money being used mostly for raters in a direct assessment of writing. This cost is tied directly to the amount of time raters spend at their job, which, in turn, is a function of the type of scale the raters use (c.f. Huot, 1990) and, the number of pieces of writing they must read. The "logical" method to reduce costs, then is to keep the number of raters and the number of writing samples as low as possible to have the process happen as quickly as possible.

Generalizability and cost: A dynamic relationship

The needs for generalizability and for cost and efficiency seemingly work against each other, since generalizability can be improved through increasing the number of observations, i.e. the number of tasks and/ or the number of raters; and cost can be reduced by lowering the number of tasks and/or the number of raters.

Several researchers have stressed the need to have a writing assessment include more than one piece of writing (e.g. Moran, Myles, and Shank, 1991; Raymond, 1982; Kegley, 1986; Breland, et al., 1987). Breland et al. (1987) conclude that such an assessment would be more valid. Therefore, it is incumbent upon researchers of direct

writing assessment to explore the possibilities of creating such an assessment, while always keeping an eye on the budget.

The relationship between generalizability and cost is a dynamic, non-linear one. The generalizability coefficient is estimated based on the variance components of raters, tasks, and other facets of the measurement situation, upon which cost also ultimately depends. In classical test theory, which can deal with only one facet of the measurement situation at a time, the relationship between the facet and the reliability coefficient is a direct one. That is, as the level of the facet increases, so does the reliability coefficient. Thus the rule of thumb as exemplified in the Spearman-Brown Prophecy Formula is increasing the number of observations will increase the reliability. However, in generalizability theory, which can recognize several facets, such a direct relationship does not necessarily exist. Increasing the generalizability coefficient no longer is a simple univariate function, but rather it becomes a combinatorial process of altering the number in each facet to produce the best possible configuration to produce the best reliability. It is actually possible to decrease the total number of observations and increase the generalizability coefficient (Sanders, Theunissen, and Baas, 1992). The entire goal then, is to find that optimal combination of facets which will produce the largest generalizability coefficient at the lowest possible cost.

Solving the puzzle -- constrained optimization

Along with the G-study option of generalizability theory, there is also a Decision Study (D-study) which allows for hypothetical situations to be considered based on existing data (Brennan, 1992). This procedure is analogous to the Spearman-Brown Prophesy Formula through which a reliability coefficient can be projected for a hypothetical number of items. Shavelson et al. (1992, 1993) and McWilliam and Ware (1994) both employed this technique to arrive at the necessary number of observations to reach a certain generalizability threshold. This technique has some limitations. It works best with facets

with a finite number of levels to maintain manageability. If a measurement situation arose with a large number of facets, the number of possible D-study scenarios would increase multiplicatively, making this technique unwieldy. Furthermore, optimization is far from guaranteed if a facet has a large number of levels, since all possible combinations could not be considered (Sanders, Theunissen, and Bass, 1991).

A second limitation enumerated by Sanders et al. (1991) is that D-studies cannot include economic considerations. Since raters cost more than items on an observation report, for instance, the cost per unit increase for one does not equal the cost per unit increase in the other. Combine that with the knowledge that raters and items probably do not contribute equally to the generalizability coefficient (Sanders et al, 1992), and the problem quickly becomes impossible for the human mind to solve unaided.

Despite their use of generalizability theory which took advantage of the dynamic interplay among aspects of the measurement situation in determining a generalizability coefficient, Shavelson et al. (1992, 1993) and McWilliam and Ware (1994) did not take the dynamic interplay of generalizability and cost factors into account when projecting the minimum number of observations necessary to reach a given generalizability coefficient threshold.

Sanders et al. (1989, 1991, 1992) proposed the use of a branch-and-bound integer programming algorithm which searches for and identifies the optimal number of levels for each facet while taking into account each facet's contribution to the generalizability and each facet's cost as well as any other practical constraint. This technique appears to be promising. This study examines the applicability of the algorithm in identification of an optimal writing assessment design.

Method

Subjects

Fifty subjects enrolled in an undergraduate educational psychology class participated in the main phase of the study. Twenty-eight percent of the sample were males and seventy-two percent were females. The sample also contained a mix of White, Asian-American, and Hispanic subjects. By class, the sample consisted of freshmen (20%), sophomores (52%), juniors (21%), seniors (5%), with the remainder unidentified. The sample had taken an average of 1.26 writing courses with a range from 0 to 3.

Procedures

Each subject read three articles -- one about instructional approaches, and two articles about performance assessments -- prior to attending the first of two 2 1/2 hour sessions. During the first session, subjects filled out a demographic questionnaire and wrote a separate 300 to 500 word essay about each of two prompts. During the second session, subjects wrote the other two prompts. The writing prompts and the orders in which they were written are provided in Figure 1. In total, they wrote an expressive piece and a persuasive piece about the instructional approaches and an expressive piece and a persuasive piece about performance assessments. Four different counterbalanced orders of the prompts were used to allow investigation of practice effects or other effects that may arise by writing the essays in a particular order.

INSERT FIGURE 1 ABOUT HERE

Scoring the Essays

Three graduate students in Educational Psychology served as raters and received some training. These raters were given the scoring rubric and discussed it; then, they scored a sample paper as a group. Using a slightly modified version of the Diederich scale

(Diederich, 1974), each rater then read all 200 pieces of writing. The seven items on the scale were summed to achieve each subject's score on each piece of writing. The scoring rubric can be found in Figure 2.

INSERT FIGURE 2 ABOUT HERE

Analyses

The results are produced in three stages. First, the variance components for each facet of the measurement situation are estimated. Next, a number of appropriate constraints are identified. Then these variance components are submitted to a branch-and-bound integer programming algorithm that will, subject to the identified constraints, produce the optimal number of each facet necessary to reach a given level of reliability and cost.

The Variance Model

In the writing assessment used in this study, three facets are involved: mode of discourse (m), writing prompt (p), and rater (r). The object of measurement is student's overall writing ability (s). In the data collection design, prompts are nested within mode (i.e., p:m) and both cross raters and students. Thus in the generalizability framework, the variance model is:

$$\sigma^2_{(X_{srpm})} = \sigma^2_s + \sigma^2_r + \sigma^2_m + \sigma^2_{p:m} + \sigma^2_{sr} + \sigma^2_{sm} + \sigma^2_{mr} + \sigma^2_{srm} + \sigma^2_{(p:m)s} + \sigma^2_{(p:m)r} + \sigma^2_{(p:m)sr}. \quad (1)$$

Figure 3 provides a Venn Diagram of these variance components. The variance components for the sample in this study were estimated through the GENOVA (Crick and Brennan, 1983) program. Based on a review of the literature on modes of discourse (e.g., Crusius, 1989), there are at most five modes in existence. Therefore, for the estimation of variance components, the universe of modes was defined as having 5 levels. For all other

facets, the universes were defined as infinite. The variance components thus estimated are shown in Table 1.

INSERT FIGURE 3 AND TABLE 1 ABOUT HERE

For all subsequent optimization analyses, the relative model of measurement was used. Thus, relative error variances were estimated through:

$$\sigma^2(\delta) = \frac{\sigma^2_{sr}}{n_r} + \frac{\sigma^2_{sm}}{n_m} + \frac{\sigma^2_{smr}}{n_r n_m} + \frac{\sigma^2_{(p:m)s}}{n_m n_p} + \frac{\sigma^2_{(p:m)sr}}{n_r n_m n_p}, \qquad (2)$$

where n_r, n_m, and n_p are the number of raters, modes, and prompts in each particular scenario respectively. The G-coefficient of interest for each scenario was thus:

$$E\rho^2 = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2(\delta)}. \qquad (3)$$

The Optimization Scenarios

A branch-and-bound integer programming algorithm, which is a linear programming technique, was employed to estimate the optimal combination of raters, prompts within modes, and modes themselves. This investigation used the solver function of Microsoft EXCEL, version 5.0, to execute the algorithm. Four different scenarios were investigated. The first optimized the problem using only psychometric constraints; the second took a relative human factor constraint into consideration; the third used a specific human factor constraint; and the fourth used specific economic constraints. For all four scenarios, the variance components from Table 1 were entered into the worksheet.

All four scenarios investigated shared a common objective function and a common set of constraints. In Scenarios 2,3, and 4, additional constraints were considered. The common problem to be solved across all constraints is:

Objective Function: Minimize $L = n_m n_{p:m} n_r$; (4)

Subject to: $$E\rho^2 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma^2(\delta)} \geq 0.8 \quad (5)$$

$n_m \leq 5$, (6)

n_m, $n_{p:m}$, and n_r are integers, (7)

and n_m, $n_{p:m}$, and $n_r \geq 1$. (8)

The objective function is to minimize the total amount of observations needed. Constraint (5) specifies the minimal acceptable level of generalizability. Constraint (6) specifies that there are no more than 5 possible modes of discourse. Constraints (7) and (8) further delimit the search to a feasibility region of positive integers.

In Scenario 1, the objective function defined in (4) subject to constraints (5) through (8) was submitted to the branch-and-bound search algorithm. The results of this search can be found in Table 2. As shown in Table 2, to attain a G-coefficient of at least 0.8, the minimum numbers are 4 modes with 2 prompts each while employing two raters to score each prompt in each mode. Based on data obtained from the sample, the average time needed to rate each prompt in each mode in this study was 0.092 hour (approximately 5 $^1/_2$ minutes). The total amount of time needed to rate the writings from n_s subjects under any given scenario is then:

$$\text{Total manhours} = n_m\, n_{p:m}\, n_r\, n_s(.092). \quad (9)$$

Applying Equation (9), the total manhours needed for Scenario 1 for 50 subjects is 73.6.

INSERT TABLE 2 ABOUT HERE

An apparent practical problem with Scenario 1 is the demand on the examinee. Based on the results of this search, to attain a G-coefficient of 0.8 or higher, each examinee must produce a total of 8 pieces of writing; 1 for each of 2 prompts within 4 different modes. Unless the 8 pieces of writing are considered a portfolio accumulated over time, producing 8 pieces of writing in a single assessment can be quite taxing on the examinee. If they were to be considered 8 pieces in a long-term portfolio, the use of 8 randomly parallel writings in a portfolio assessment would be problematic as the added value of each piece to the portfolio is not optimal.

A better solution might be one in which the burden of reliability is shifted away from the demand on the examinee to a demand on ratings per piece of writing. In Scenario 2, a new constraint was added to shift this demand to ratings. Specifically, the following constraint was added to constraints (5) through (8):

$$n_r \geq n_m\, n_{p:m}. \quad (10)$$

The results of this search can be found in Table 2. To attain a G-coefficient of at least 0.8 while minimizing the burden on the examinee, the minimal design is one in which each examinee responds to 4 different prompts in a single mode of discourse. Each piece of writing needs to be rated by 5 raters. Under this scenario, the total number of writings from each examinee is only four. However, the total amount of manhours needed for the rating of 50 subjects increases to 92 manhours.

In Scenario 3, a compromise between Scenarios 1 and 2 was investigated. Specifically, instead of constraint (10), in which the burden is specifically shifted to the number of ratings and thus total manhours, the following constraint was used:

$$n_m\, n_{p:m} \leq 6. \qquad (11)$$

Through constraint (11), the total number of writings needed from each examinee is limited to 6 or less. The result of the search for this scenario is shown in Table 2. As can be seen, under this scenario, each examinee must produce 6 pieces of writing in a single mode. On the other hand, only 3 raters are needed for each piece to attain a G-coefficient of 0.8 or higher. The total manhours for 50 subjects in this case is 82.8.

Scenario 4 investigated the cost factor. Specifically, the cost of rating in the form of total number of manhours was examined. Assuming that the cost per manhour is fixed, the total cost of rating the writings from n_s subjects is as defined in Equation (9). For this scenario, n_s was defined as 50. Based on Equation (9), the following constraint was used in place of constraint (11):

$$n_m\, n_{p:m}\, n_r\, (50)(.092) \leq 60 \text{ manhours}. \qquad (12)$$

Combining constraints (5) through (8) with constraint (12) produced no feasible solution. Relaxing constraint (12) to

$$n_m\, n_{p:m}\, n_r\, (50)(.092) \leq 70 \text{ manhours}. \qquad (13)$$

also did not lead to a feasible solution. In other words, it is not possible to expend less than 70 manhours of rating activities to rate the writings used in this study for 50 subjects and still maintain a minimum G-coefficient of 0.8.

Discussion

This investigation successfully demonstrates the advantages that the optimization algorithm has over other previously used techniques. Some cautions need to be issued, however, regarding the actual estimates presented here. First, this study reports a generalizability coefficient of 0.75, which is quite high in comparison with other investigations. Shavelson, Baxter, and Gao (1993) report G-coefficients at approximately 0.2 or 0.3. Koretz, Klein, McCaffrey, and Stecher (1994) report reliability coefficients ranging from 0.3 to nearly 0.8. There are a number of reasons why the estimate in this study is higher than those previously reported. The variance component for mode was estimated as negative and was therefore set to zero. Mode was hypothesized to contribute to the relative error variance, yet it did not. This could have occurred because the subjects wrote the four pieces about topics in which they had a common background, since the topics were taken from topics being discussed in the introductory educational psychology class from which the sample was drawn.

Another caution is that the estimates for the number of tasks are considerably lower than other estimates. Shavelson, Baxter, and Gao (1993) projected between 8 and 23 tasks would be necessary to achieve a G-coefficient of 0.8 or higher. The present investigation has 8 tasks as the maximum necessary. This difference could be due to the higher reliability estimate of the actual data used to make the projections. Regardless of these limitations in the actual numbers involved, the efficacy and preferability of the branch-and-bound algorithm have been demonstrated.

There are several ways in which to add economic constraints to the problem, some of which have been demonstrated here. Sanders, Theunissen and Bass (1991) used cost in dollars to constrain the search. In this investigation, number of total manhours was employed. Using total manhours provides some flexibility. As expressed in Equation (9),

total manhours is a function of the amount of time it takes for one reading of one piece of writing for one subject. In this context, n_r does not represent the number of raters, per se, but rather the number of ratings per piece per student. The actual number of raters needed is up to the assessment developers; n_r expresses the minimum number of people necessary. That is how total manhours adds flexibility: the developers could choose to hire as many people as they wish. If a deadline required a rapid scoring, then the total manhours could be spread out across many people rather than across time.

Making projections about what would be necessary to increase the reliability of an assessment is not always a straightforward process. For unidimensional measurements, such as a multiple-choice test, the Spearman-Brown Prophesy Formula works well because the relationship between the number of items and the estimated reliability is direct. In that case, it is a rather straightforward process. When a multidimensional assessment situation exists, however, the relationship between the levels of the facets and the estimated reliability is much more complicated. Since not every facet contributes equally to the total variance, and thus to the estimated reliability, there is no simple representation like Spearman-Brown with which to make projections. D-studies are currently employed to make such projections, though they do not ensure that the optimal combination of levels of facets has been found. Further, they can become unwieldy as the number of facets and the number of levels increase. The constrained branch-and-bound integer programming technique ensures both an optimal solution, if one exists, and manageability.

In a unidimensional assessment, the effects of cost are fixed in the sense that cost only arises from the one dimension, i. e. item development, and therefore cannot be altered. When considering a multidimensional assessment, each facet conceivably could have a different cost associated with it. So not only do the different facets contribute unequally to the error variance, but they also contribute unequally to the total cost of the assessment. Such considerations are beyond the scope of a D-study. Only through the use of the constrained optimization technique can both reliability and cost be considered

simultaneously. Finally, other meaningful considerations such as the demands on the examinee can be taken into account as well. It is this ability which makes the branch-and-bound integer programming approach preferable to Spearman-Brown or D-studies in a multidimensional measurement situation.

References

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., and Rock, D. A. (1987). *Assessing writing skills*. New York: College Entrance Examination Board.

Brennan, R. L. (1992). *Elements of Generalizability Theory*. Iowa City: ACT.

Crick, J. E., and Brennan, R. L. (1982). *GENOVA: A generalized analysis of variance system* (FORTRAN IV computer program and manual). Iowa City: ACT.

Crusius, T. W. (1989). *Discourse: A critique and synthesis of major theories*. New York: Modern Language Association of America.

Diederich, P. B. (1974). *Measuring Growth in English*. Urbana, IL: National Council of Teachers of English.

Engelhard, G., Gordon, B., and Gabrielson, S. (1992). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. *Research in the Teaching of English, 26 (3)*, 315- 336.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60 (2)*, 237-263.

Kegley, P. H. (1986). The effect of mode of discourse on student writing performance: Implications for policy. *Educational Evaluation and Policy Analysis, 8 (2)*, 147-154.

Koretz, D., Klein, S., McCaffrey, D., and Stecher, B. (1994). *Interim report: The reliability of the Vermont portfolio scores in the 1992-93 school year*. (RAND/ RP - 260). Santa Monica, CA: RAND.

Koretz, D., Stecher, B., Klein, S., and McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13 (3)*, 5-16.

Koretz, D., Stecher, B., Klein, S., McCaffrey, D., and Deibert, E. (1994). *Can portfolios assess student performance and influence instruc n?*. (RAND/ RP-259). Santa Monica, CA: RAND.

Linn, R. L., Baker, E. L., and Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20 (8)*, 15-21.

McWilliam, R. A., and Ware, W. B. (1994). The reliability of observations of young children's engagement: An application of generalizability theory. *Journal of Early Intervention, 18 (1)*, 34-47.

Moran, M. R., Myles, B. S., and Shank, M. S. (1991). Variables in eliciting writing samples. *Educational Measurement: Issues and Practice, 10 (3)*, 23-26.

Prater, D. and Padia, W. (1983). Effects of mode of discourse on writing performance in grades four and six. *Research in the Teaching of English, 17 (2)*, 127-134.

Quellmalz, E. S., Capell, F. J., and Chou, C. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement, 19*, 241-258.

Raymond, J. C. (1982). What we don't know about the evaluation of writing. *College Composition and Communication, 33(4)*, 56-59.

Sanders, P. F., Theunissen, T. J. J. M., and Baas, S. M. (1989). Minimizing the number of observations: A generalization of the Spearman-Brown formula. *Psychometrika, 54* , 587-598.

Sanders, P. F., Theunissen, T. J. J. M., and Baas, S. M. (1991). Maximizing the coefficient of generalizability under the constraint of limited resources. *Psychometrika, 56 (1)*, 87-96.

Sanders, P. F., Theunissen, T. J. J. M., and Baas, S. M. (1992). The optimization of decision studies. In M. Wilson (Ed.) *Objective Measurement: Theory into Practice (Vol. 1)*. Norwood, NJ: Ablex.

Shavelson, R. J. and Baxter, G. P. (1992). What we've learned about assessing hands-on science. *Educational Leadership, 49 (8)*, 20-25.

Shavelson, R. J., Baxter, G. P., and Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30 (3)*, 215-232.

White, E. M. (1986). Pitfalls in the testing of writing. In K. L. Greenberg, H. S. Wiener, and R. A. Donovan (Eds.), *Writing Assessment: Issues and strategies*. New York: Longman.

White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance, 2nd ed.* San Francisco, CA: Jossey-Bass Publishers.

Table 1

Source of Variation	Estimated Variance
Subject (s)	5.8275728
Mode (m)	0*
Prompt : mode (p:m)	0*
Rater (r)	5.6756912
sm	0*
s(p:m)	2.6025238
sr	0.6714422
smr	0.3008503
sr(p:m)	11.8791415

* Negative variance components were set equal to zero, following Brennan (1992).

Table 2

	Actual	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Mode	2	4	1	1	4
Prompt: Mode	2	2	4	6	2
Rater	3	2	5	3	2
Obj. Function	12	16	20	18	16
Manhours	55.2	73.6	92	82.8	73.6
G Coefficient	0.752713326	0.801739412	0.801964784	0.804316069	0.801739412

Figure 1

The Writing Prompts and the Orders

in Which They Were Written

Based on Article #29, "Optimizing the Instructional Moment: Guide to Using Socratic, Didactic, Inquiry, and Discovery Methods," organize and write a 300-500 word response to this question.

QUESTION I. A. : Some of these methods seem to be better or more appealing than the others. Which one or ones do you think are the best methods? Argue the point, and explain your reasons for choosing as you did.

Based on Article #41, "Creating Tests Worth Taking," AND Article #42, "Innovation or Enervation?: Performance Assessment in Perspective," organize and write a 300-500 word response to the following question.

QUESTION P. E.: We have all taken standardized, multiple choice tests. Describe an experience you remember with one such test (SAT, ACT, PSAT, NTE, GRE, Achievement Tests). How did you feel? What did you think? Based on those experiences, what's your opinion about such tests?

Based on Article #29, "Optimizing the Instructional Moment: Guide to Using Socratic, Didactic, Inquiry, and Discovery Methods," organize and write a 300-500 word response to the following question.

QUESTION I.E. : The didactic method is used quite heavily in college classrooms, especially in large lecture classes. Describe what such a classroom is like. How do you feel as a student in that type of classroom? Can you relate any particular stories about such classrooms, either from college or your earlier educational experiences?

Based on Article #41, "Creating Tests Worth Taking," AND Article #42, "Innovation or Enervation?: Performance Assessment in Perspective," organize and write a 300-500 word response to the following question.

QUESTION P. A.: Argue for or against performance assessments. Should they replace multiple choice standardized tests? Why or Why not?

The four orders were:

ORDER 1: Instructional method (expressive); Instructional method (persuasive); Performance assessment (persuasive); performance assessment (expressive).

ORDER 2: Performance assessment (expressive); Performance assessment (persuasive); Instructional method (persuasive); Instructional method (expressive).

ORDER 3: Performance assessment (persuasive); Instructional method (expressive); Performance assessment (expressive); Instructional method (persuasive).

ORDER 4: Instructional method (persuasive); Performance assessment (expressive); Instructional method (expressive); Performance assessment (persuasive).

Figure 2

The Scoring Scale*

	LOW		MODERATE		HIGH
IDEAS	1	2	3	4	5
ORGANIZATION	1	2	3	4	5
FLAVOR	1	2	3	4	5
WORDING	1	2	3	4	5
USAGE	1	2	3	4	5
MECHANICS	1	2	3	4	5
SPELLING	1	2	3	4	5

* See Diederich (1974) for a complete explication.

Figure 3